Unc
SECURI

AD-A199 706

(1)

DOCUMENTATION PAGE

1a. REP Unc	1b. RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for Public Release; Distribution Unlimited
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) AFGL-TR-88-0267	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION Air Force Geophysics Laboratory	6b OFFICE SYMBOL (If applicable) LIS
6c. ADDRESS (City, State, and ZIP Code) Hanscom AFB Massachusetts 01731-5000	7a. NAME OF MONITORING ORGANIZATION 7b ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING / SPONSORING ORGANIZATION	8b OFFICE SYMBOL (If applicable)
8c. ADDRESS (City, State, and ZIP Code)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO	PROJECT NO.	TASK NO	WORK UNIT ACCESSION NO
62101F	4643	09	03

11 TITLE (Include Security Classification)
Maximum Entropy Calculations on a Discrete Probability Space

12 PERSONAL AUTHOR(S)
P.F. Fougere

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Reprint	FROM ___ TO ___	1988 September 30	30

16 SUPPLEMENTARY NOTATION
Reprinted from Maximum-Entropy and Bayesian Methods in Science and Engineering (Vol 1) 205-234

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP

18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Maxtent, Maximum Entropy, Discrete Probability Space, Wolf's Dice Data

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

The Maximum Entropy Principle

In a remarkable series of papers beginning in 1957, E. T. Jaynes (1957) began a revolution in inductive thinking with his principle of maximum entropy. He defined probability as a degree of plausibility, a much more general and useful definition than the frequentist definition as the limit of the ratio of two frequencies in some imaginary experiment. He then used Shannon's definition of entropy and stated that in any situation in which we have incomplete information, the probability assignment which expresses all known information and is maximally non-committal with respect to all unknown information is that unique probability distribution with maximum entropy (ME). It is also a combinatorial theorem that the unique ME probability distribution is the one which can be realized in the

(Cont'd)

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION Unclassified
22a NAME OF RESPONSIBLE INDIVIDUAL P.F. Fougere	22b. TELEPHONE (Include Area Code) (617) 377-2692
	22c. OFFICE SYMBOL LIS

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

Cont of Block 19:

greatest number of ways. The ME principle also provides the fairest description of our state of knowledge. When further information is obtained, if that information is pertinent then a new ME calculation can be performed with a consequent reduction in entropy and an increase in our total information. It must be emphasized that the ME solution is not necessarily the "correct" solution; it is simply the best that can be done with whatever data are available. There is no one "correct solution", but an infinity of possible solutions. These ideas will now be made quite concrete and expressed mathematically.

MAXIMUM ENTROPY CALCULATIONS ON A DISCRETE PROBABILITY SPACE

P. F. Fougere
AFGL/LIS
Hanscom AFB, Bedford, MA

To Ed Jaynes, who started it 30 years ago and whose clarity of exposition is an inspiration to us all.

I. The Maximum Entropy Principle

In a remarkable series of papers beginning in 1957, E. T. Jaynes (1957) began a revolution in inductive thinking with his principle of maximum entropy. He defined probability as a degree of plausibility, a much more general and useful definition than the frequentist definition as the limit of the ratio of two frequencies in some imaginary experiment. He then used Shannon's definition of entropy and stated that in any situation in which we have incomplete information, the probability assignment which expresses all known information and is maximally non-committal with respect to all unknown information is that unique probability distribution with maximum entropy (ME). It is also a combinatorial theorem that the unique ME probability distribution is the one which can be realized in the greatest number of ways. The ME principle also provides the fairest description of our state of knowledge. When further information is obtained, if that information is pertinent then a new ME calculation can be performed with a consequent reduction in entropy and an increase in our total information. It must be emphasized that the ME solution is not necessarily the "correct" solution; it is simply the best that can be done with whatever data are available. There is no one "correct solution", but an infinity of possible solutions. These ideas will now be made quite concrete and expressed mathematically.

(a) Discrete Probability Space.

We have n propositions or statements, S_1, S_2 . . .S_n, each of which can be assigned a probability p_i, i = 1,n. The number p_i runs from zero when our information tells us that S_i is not true to one when we assume that S_i is true. In the case of a die, S_i might be the proposition that on the next throw of the die face i will be up. If the die has not yet been cast then our belief that face i will come up next is described by assigning a number to p_i. If the die were perfectly symmetric and thrown in a fair way, making no attempt to favor any face, then every face would be equally likely to occur and then since one of them must occur, the probability of the statement "some i will occur" is 1. Thus the probabilities would each be set to 1/n; in the case of a die (p_i = 1/6, i=1,6). This is a simple

G. J. Erickson and C. R. Smith (eds.),
Maximum Entropy and Bayesian Methods in Science and Engineering (Vol. 1), 205–234

expression of Laplace's "principle of insufficient reason " which has been attacked by many but has never been replaced. It is essentially a symmetry principle. If the mechanism of selecting a number at random from the possible set of n is symmetric with respect to all members of the set then the probability of each is 1/n. There are many practical realizations of this mechanism of selection. All of the resulting problems are isomorphic and all can be solved in precisely the same way.

1. There are n distinguishable but otherwise identical objects numbered 1, 2, ...n in an opaque container. An experiment consists of selecting an object, noting its number and replacing the object in the container.

2. A roulette wheel containing 36 numbered slots is spun and a small ball is set in motion in the opposite direction. When both wheel and ball slow down sufficiently the ball drops into one of the slots. The number is recorded.

3. An ordinary 6 sided die is thrown. The number of spots facing up is recorded.

4. A deck of 52 playing cards is shuffled face down. A card is selected and its value noted.

Note that there may be bias introduced either accidently or deliberately (to cheat) in any of these games. But also note that if the bias (a favoring of any outcome over the others) becomes large enough, the players of the game will almost certainly notice, with retribution to the perpetrator soon to follow. Cheats at poker, craps (dice) and roulette have often met an untimely end!

We will soon see that the ME method is admirably suited to detecting such biases, even very tiny ones. Every time a correctly calculated ME probability distribution fails to reproduce an observed frequency distribution accurately enough, the conclusion can be drawn that a bias which has not yet been taken into account is operating. In just this way was quantum mechanics discovered!

The principle of insufficient reason will be derived as the maximum entropy assignment: given only an enumeration of the possibilities and normalization:

$$\Sigma\, p_i = 1, \qquad (1)$$

and nothing else.

Throughout this article, sums on i will always run from 1 to n, and for simplicity of notation the limits will not be typed. The ME probability distribution given only the above information is (p_i = 1/n, i=1, 2 ...n). This statement will be proved in Section b. This

expresses exactly the known information and nothing more. Any subsequent information which is provided, for example: "the die is not symmetric", will lower the entropy and change the probabilities accordingly.

(b) Entropy.

In his wonderful little book on information theory Shannon (1948) first set forth the axioms or elementary desiderata of consistency as follows: if S is the measure of information or uncertainty and p_i = probability of the i'th outcome:

1. $S = S(p_1, p_2, \ldots p_n)$

The information depends upon the entire set of probabilities.

2. If all p_i are equal then S is a monotone increasing function of n. With more possibilities to choose from the information in a choice is greater.

3. S is additive for compound independent events. If events A and B are independent, S (AB) = S(A) + S(B). The information contained in the statement "it is raining and today is Tuesday" is exactly equal to the information contained in the statement "it is raining" plus the information contained in the statment "today is Tuesday".

4. S does not depend upon how the problem is setup. See Figure 1.

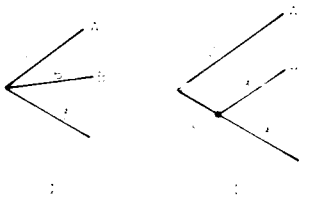

Figure 1. Two sets of probability assignments. In 1a there are three events A, B, C with probabilities 1/2, 1/6, 1/3 respectively. In 1b the final state A, B, C is reached via an intermediate state D with probability 1/2. The information in both diagrams at stage A, B, C must be the same.

The information in the probability assignment A = 1/2, B = 1/6, C = 1/3 in Figure 1a must be the same as that in Figure 1b where we have used the intermediate point D.

Shannon then proved [see also Tribus (1961, 1969)] that this measure of information has the form:

$$H = - K \Sigma p_i \log p_i \qquad (2)$$

and furthermore that this functional form is unique: it is the only form capable of satisfying the four axioms. The constant K is merely a scale factor and the base of the logarithm is arbitrary; for convenience the constant K is set to 1 and the base of the logarithm is taken to be natural. Thus we have:

$$H = - \Sigma p_i \ln p_i \qquad (3)$$

Since the p_i are all in [0,1], $H > 0$, if we agree that $0 \ln 0 = 0$, (a proposition which has zero probability conveys no information). As an elementary exercise let us prove that the probability assignment with maximum entropy is one with $p_i = 1/n$.

We have $\Sigma p_i = 1, \qquad H = - \Sigma p_i \ln p_i \qquad (4)$

Form the expression $Q = - \Sigma p_i \ln p_i + \lambda (\Sigma p_i - 1)$

Where λ is a Lagrange multiplier used to enforce normalization.

Now differentiate with respect to p_1:

$$\frac{\partial Q}{\partial p_j} = - (\ln p_j + 1) + \lambda = 0$$

thus $\ln p_j = \lambda - 1$

then $p_j = \exp (\lambda - 1) \qquad (5)$

But this is independent of j. Thus all p_j are equal and by normalization they sum to 1; therefore $p_j = 1/n$, j=1,n. Thus with only an enumeration of the possibilities which are exhaustive (one must occur) and exclusive (only one can occur) and normalization, the probability assignment which maximizes the entropy brings us back to Laplace's principle of insufficient reason. Any further information would change the probabilities and lower the entropy. We do not need Laplace's principle of insufficient reason; entropy maximization subject only to normalization produces Laplace's principle as a theorem or result.

(c) Maximum Entropy Formalism.

Since we will be maximizing entropy under a variety of constraints, it is helpful to have "cookbook recipe" or a "crank to turn".

In addition to normalization (Eq. 1) we may have M constraints in the form of expectation values or averages in the form:

$$\Sigma\, p_i\, f_m(x_i) = \langle f_m \rangle = F_m, \quad m = 1, 2 \ldots M \tag{6}$$

We use the calculus of variations now and take variations of our important equations 4 and 6 to get:

$$\begin{aligned} \delta H &= -\Sigma\,(1 + \ln p_i)\,\delta p_i = 0 \\ (\lambda_0 - 1)\,\Sigma\,\delta p_i &= 0 \\ \sum_m \lambda_m \sum_i f_m(x_i)\,\delta p_i &= 0 \end{aligned} \tag{7}$$

$\lambda_0, \lambda_1 \ldots \lambda_M$ are, of course, Lagrange multipliers. Now add the three equations and factor δp_i:

$$\sum_i \left[1 + \ln p_i + \lambda_0 - 1 + \sum_m \lambda_m f_m(x_i) \right] \delta p_i = 0 \tag{8}$$

For any arbitrary variation, δp_i, the expression in brackets must vanish for every value of i. Solving for $\ln p_i$ we get

$$\ln p_i = -\lambda_0 - \sum_m \lambda_m f_m(x_i)$$

Thus

$$p_i = \exp\left[-\lambda_0 - \sum_m \lambda_m f_m(x_i) \right] \tag{9}$$

Now for normalization we have that

$$\Sigma\, p_i = 1 = \sum_i \exp\left[-\lambda_0 - \sum_m \lambda_m f_m(x_i) \right] \tag{10}$$

Solving for $\exp(\lambda_0)$, which we call the partition function Z:

$$Z = \exp(\lambda_0) = \sum_i \exp\left[-\sum_m \lambda_m f_m(x_i) \right] \tag{11}$$

Taking logs of both sides

$$\lambda_0 = \ln \sum_i \exp\left[-\sum_m \lambda_m f_m(x_i) \right] \tag{12}$$

Thus λ_o is the log of the partition function Z; for reasons which will become clear immediately we call λ_o the potential function.

Now differentiate λ_o with respect to r

$$\frac{\partial \lambda_0}{\partial \lambda_r} = \frac{-\sum_i f_r(x_i) \exp\left[-\sum_m \lambda_m f_m(x_i)\right]}{\sum_i \exp\left[-\sum_m \lambda_m f_m(x_i)\right]} \quad (13)$$

Multiply numerator and denominator by $\exp(-\lambda_o)$

Then

$$\frac{\partial \lambda_0}{\partial \lambda_r} = \frac{-\sum_i f_r(x_i) \exp\left[-\lambda_0 - \sum_m \lambda_m f_m(x_i)\right]}{\sum_i \exp\left[-\lambda_0 - \sum_m \lambda_m f_m(x_i)\right]} \quad (14)$$

Now notice from Eq. 9 that the exponential of the bracketed term in numerator and denominator is *just the probability* p_i. Thus

$$\frac{\partial \lambda_0}{\partial \lambda_r} = \frac{-\sum_i f_r(x_i) p_i}{\sum_i p_i} = -\langle f_r \rangle \quad (15)$$

We now see that λ_o is called the potential function because the constraints are given as derivatives of λ_o with respect to all the other λ's.

For convenience we now summarize the important formulas:

$$Z = \sum_i \exp\left[-\sum_m \lambda_m f_m(x_i)\right]$$

$$\frac{\partial \ln Z}{\partial \lambda_m} = -\langle f_m \rangle \quad (16)$$

$$p_i = \exp\left[-\sum \lambda_m f_m(x_i)\right] / Z$$

We have exactly one Lagrange Multiplier λi for each constraint and we determine the set of λ's by solving the MXM set of equations

$$\frac{\partial}{\partial \lambda_m} \ln Z (\lambda_1, \lambda_2, \ldots\ldots, \lambda_M) = - F_m \qquad (17)$$

Finally the probabilities are given by:

$$p_i = 1/Z \exp \left[- \lambda_1 f_1(x_i) - \lambda_2 f_2 (x_i) \ldots - \lambda_M f_M (x_i) \right] \qquad (18)$$

We can seee immediately that $\sum p_i = Z/Z=1$ and thus the formalism automatically produces a normalized set of p_i.

II. Wolf's Dice Data

To make the foregoing ideas as concrete as possible we will now examine in detail a remarkable series of experiments performed about 100 years ago by the Swiss scientist Rudolf Wolf who is known well for his work on sunspots. One of the experiments, reported by Czuber(1908), consisted of throwing a pair of dice, one red, the "ROTER WÜRFEL" and the other white, the "WEISSER WÜRFEL", a total of 20,000 times. The dice were thrown carefully in such a way as to avoid as much as possible introducing any bias, any artificial favoring of any of the 6 sides. Evidently (as we shall see) the dice were made using ordinary care but not extraordinary care - they were in fact quite noticeably biased.

Ed Jaynes has written extensively on dice in general and on Wolf's dice data in particular in no less than four publications (1963a, 1978, 1979, 1982). I would urge the reader to look up and read this exciting scientific saga. I freely acknowledge my deep indebtedness to Ed Jaynes for my inspiration in writing this paper but of course any mistakes which I may have made in interpretation, emphasis, algebra or arithmetic are mine alone.

Table I lists the totals obtained by Wolf for the 36 distinct possibilities - that is: white 1 red 1; white 1 red 2; . . . up to white 6 red 6.

Table I Wolf's Dice Data:

		Weisser Würfel						RM	RF
	NR.	1	2	3	4	5	6		
Roter Würfel	1	547	587	500	462	621	690	3407	0.17035
	2	609	655	497	535	651	684	3631	0.18155
	3	514	540	468	438	587	629	3176	0.15880
	4	462	507	414	413	509	511	2916	0.14580
	5	551	562	499	506	658	672	3448	0.17240
	6	563	598	519	487	609	646	3422	0.17110
WM		3246	3449	2897	2841	3635	3932	20,000	
WF		.16230	.17245	.14485	.14205	.18175	.19660		1.0

RM and WM are the red and white marginals, respectively.
RF and WF are the red and white relative frequencies, respectively.

Since there is no evidence for and no reason to expect that the two dice were correlated, the results for the white die are independent of those for the red die, and Table I also lists the white marginals, the total number of times that the white die came up a given number of spots independent of which red spot was showing. Similary the red marginals are listed. It can be seen at once that the dice were indeed biased; for example W6 appeared 3932 times, almost 600 times more than expected if the die were fair; W4 appears only 2841 times, 492 times less than expected. The relative frequencies given in Table I are just the marginals divided by 20,000.

a. The White Die

Let us now, following Ed Jaynes, try to account for some of the discrepancies or biases using ME. At this point, it is important to know what a conventional playing "die" is. It is a solid cubical object, made of a machineable substance such as ivory. Hemispherical depressions or excavations (spots) are made symmetrically in each face, with the number of spots on opposite faces totaling 7. The spots are painted in a contrasting color. Thus 1 is opposite 6, 2 opposite 5 and 3 opposite 4. If face 6 is "up" and face 2 is visible, then face 4 is to the right of face 2. The reader's intuition will be aided by actually examining a real die.

1. One constraint. The most obvious physical asymmetry is now apparent. Whereas six spots are removed from face 6 only one is removed from face 1 and thus the center of gravity of the die is shifted very slightly toward the 1 face. Similarly the 2 and 3 faces are slightly heavier than their opposites 5 and 4 respectively. Quantitatively, the center of gravity will be shifted toward the "3" face by small distance ϵ corresponding to a one-spot discrepancy. Similarly the center of gravity will be shifted toward the "2" face by 3ϵ and towards the "1" face by 5ϵ. Thus the spot frequencies should

be shifted proportionally (frequency shift = α times center of gravity shift = $\alpha\epsilon$). Then the spot frequencies should vary linearly with i:

$$g_i = 1/6 + \alpha\epsilon\, f_1(i) \tag{19}$$

Where $f_1(i) = i - 3.5$.

Thus the expected number of spots would be shifted to (all of the sums on i will now run from 1 to 6.)

$$\langle i \rangle = \Sigma\, i\, g_i = 3.5 + 17.5\, \alpha\epsilon \tag{20}$$

or the function $f_1(i)$ has a non-zero expectation:

$$\langle f_1 \rangle = 17.5\, \alpha\epsilon \tag{21}$$

We note by calculating from Table I that the average number of spots showing on the white die was 3.5983. This was larger than 3.5 as expected on the physical grounds just discussed and not equal 3.5 as would have been expected from a fair die. Let us use this one piece of information as a constraint and find the six p_i's which yield maximum entropy. The complete statement of the problem at this stage is: We are given 1: an enumeration of the possibilities, namely i = 1,2,3,4,5,6 and 2 : $\langle i \rangle$ = A and nothing else. It is thus simpler to use $h(x_i) = i$ as constraint function, rather than $f_1(x_i) = i - 3.5$, because we are given the average value of h = A. The ME equations become:

$$Z = \Sigma \exp \lambda h(x_i); \quad h(x_i) = i;$$
$$\Sigma\, p_i\, h(x_i) = \Sigma\, i\, p_i = A \tag{22}$$

Let $y = \exp(\lambda)$

$$Z = \Sigma\, (\exp(\lambda))^i$$
$$= \Sigma\, y^i = y + y^2 + y^3 + y^4 + y^5 + y^6 \tag{23}$$

$$\frac{\partial \ln Z}{\partial \lambda} = y/Z\left[1 + 2y + 3y^2 + 4y^3 + 5y^4 + 6y^5\right] = A \tag{24}$$

Expanding and simplifying we get:

$$(1 - A) + (2 - A)y + (3 - A)y^2 + (4 - A)y^3$$
$$+ (5 - A)y^4 + (6 - A)y^5 = 0 \tag{25}$$

This 5'th degree equation has one real root; Table II gives the value of the real root y versus the average A. Here we have used the IMSL subroutine "ZPOLY".

Table II. Root of Eq. 20 (y) versus average value (A).

A	y	A	y	A	y	A	y	A	y
1.0	.000000	2.0	.532820	3.0	.839769	4.0	1.190804	5.0	1.876805
1.1	.090912	2.1	.565943	3.1	.870434	4.1	1.235307	5.1	2.006740
1.2	.166756	2.2	.597991	3.2	.901644	4.2	1.282800	5.2	2.164185
1.3	.231313	2.3	.629215	3.3	.933540	4.3	1.333821	5.3	2.360807
1.4	.287438	2.4	.659827	3.4	.966271	4.4	1.389030	5.4	2.616096
1.5	.337239	2.5	.690010	3.5	1.000000	4.5	1.449254	5.5	2.965257
1.6	.382249	2.6	.719927	3.6	1.034906	4.6	1.515549	5.6	3.479017
1.7	.423584	2.7	.749726	3.7	1.071191	4.7	1.589282	5.7	4.323151
1.8	.462068	2.8	.779545	3.8	1.109085	4.8	1.672267	5.8	5.996777
1.9	.498321	2.9	.809516	3.9	1.148853	4.9	1.766964	5.9	10.999661

For Wolf's white die, we had A = 3.5983 giving y = 1.034392, Z=6.76292. The ME probabilities are $p_i = y^i/Z$ and are given in Table III.

Table III. Wolf's dice data with one constraint (white die)

i	g_i	p_i	$\Delta_i = g_i - p_i$	C_i
1	0.16230	0.15294	0.0094	11.46
2	0.17245	0.15818	0.0143	25.75
3	0.14485	0.16361	- 0.0188	43.02
4	0.14205	0.16922	- 0.0272	87.25
5	0.18175	0.17502	0.0067	5.18
6	0.19660	0.18103	0.0156	26.78
				199.43

g_i are the relative frequencies (WF) from Table I.

p_i are the ME probabilities based on the constraint: A = <i> = 3.5983.

$C_i = 20{,}000_2(g_i-p_i)^2/p_i$ = Partial contribution to Chi^2. The critical value: $Chi^2_c(0.05) = 9.49$ on 4 degrees of freedom. The concept of degrees of freedom will be discussed later.

Examining Table III carefuliy we see that the deviations, $\Delta_i = g_i - p_i$ between observed relative frequencies, g_i, and ME probabilities, p_i, are negative for faces 3 and 4 and positive for faces 1, 2, 5, 6 and the C_i tell us that these deviations are highly significant. This does not mean that ME has failed but that there is a further physical constraint. At this point in Jaynes' paper he again demonstrates his genius as a practical working physicist, who as Enrico Fermi did, now delights in going into the machine shop to make things work. Jaynes explains to us just how to turn a lump of ivory into as perfect a cube as possible. A milling machine used by an expert would have no

trouble in cutting 5 sides of the die all accurately plane with all angles accurately 90° and the top face accurately square. But then the die would have to be removed from the machine and turned upside down to finish to final face. It would be extremely difficult to ajust the work table height so that the final dimension is exactly equal to the other two: The result of the difficulty would be a die which is either: (i) slightly "oblate" with one dimension shorter than the other two or (ii) slightly prolate with one dimension slightly greater than the other two. Of course either type of imperfection would constitute a "constraint" and would change the relative frequencies.

2. Two Constraints. We can now see, quite clearly, that the white die must have been prolate with the 3 - 4 dimension being slightly greater than the 1 - 6 and 2 - 5 dimensions! See Figure 2 for an exaggerated sketch of a prolate die. Such a die is more likely to fall "flat" with a 1, 2, 5 or 6 showing and thus frequencies of 3 and 4 spots would be lower than the frequencies of 1, 2, 5 or 6 spots.

Figure 2. A prolate die with the 3-4 (top - bottom) dimension B slightly larger than the other two equal dimensions A (1-6 and 2-5).

Suppose that the 3 - 4 dimension were greater than the other two by an amount δ. This would increase the frequencies g_1, g_2, g_5, g_6 by a proportional amount: $\beta\delta$ and decrease the frequencies g_3 and g_4 by an amount $2\beta\delta$ (this preserves normalization).

Thus we now define a new constraint function:

$$f_2(i) = 1, 1, -2, -2, 1, 1, \qquad (26)$$

and we find $$\langle f_2 \rangle = \Sigma\, g_i f_2(i) = g_1 + g_2 - 2(g_3 + g_4) + g_5 + g_6 = 0.1393 \qquad (27)$$

from Wolf's data on the white die given in Table I. We will have two Lagrange multipliers and the partition function Z will now be:

$$Z(\lambda_1, \lambda_2) = \Sigma \exp\left[- \lambda_1 f_1(i) - \lambda_2 f_2(i) \right] \qquad (28)$$

where $f_1(i) = i - 3.5$ from Eq. 19 and $f_2(i)$ is given in Eq. 26.

letting $x = \exp(-\lambda_1)$; $y = \exp(-\lambda_2)$

Then $Z(\lambda_1,\lambda_2) = x^{-5/2}\, y\, (1 + x + x^2y^{-3} + x^3y^{-3} + x^4 + x^5)$

We now have two constraint equations:

$$Z F_1 - x\frac{\partial Z}{\partial x} = 0, \quad Z F_2 - y\frac{\partial Z}{\partial y} = 0 \tag{29}$$

These yield two coupled equations in x and y:

$$(2F_1+5) + (2F_1+3)\, x + (2F_1+1)\, x^2y^{-3} + (2F_1-1)\, x^3y^{-3} + (2F_1-3)\, x^4 + (2F_1-5)\, x^5 = 0 \tag{30}$$

and $$(F_2-1)\,(1+x+x^4+x^5) + (F_2+2)\,(x^2+x^3)\, y^{-3} = 0$$

The IMSL library now comes to our aid with a very nice subroutine ZSPOW, which solves n simultaneous non-linear equations in n unknowns. For x and y we get 1.03223 and 1.07442 and the resulting ME probabilities are given in Table IV. $Z = 6.08530\, x^{-2.5}\, y$.

Table IV Wolf's dice data with two constraints (white die)

i	g_i	p_i	$\Delta_i = g_i - p_i$	C_i
1	0.16230	0.16433	-0.00203	0.50
2	0.17245	0.16963	0.00282	0.94
3	0.14485	0.14117	0.00368	1.91
4	0.14205	0.14573	-0.00368	1.85
5	0.18175	0.18656	-0.00481	2.48
6	0.19660	0.19258	0.00402	1.68
				9.37

See the footnote for Table III. Chi^2_c (0.05) on 3 degrees of freedom is 7.81.

Table IV agrees with Ed Jaynes' results except that he used 5 degrees of freedom and the critical value of Chi^2 at the 5% level is 11.07. He thus concluded that "there is now no statistically significant evidence for any further imperfection. . ..". In a later paper Jaynes (1979) discusses the number of degrees of freedom he should have been using and concludes unequivocally that the correct formulation is:

$$df = n - 1 - m$$

where df = number of degrees of freedom in Chi^2, n = number of possibilities (= 6 for a die) and m = number of constraints. We subtract one more for normalization. Simply put, the number of degrees of freedom is the number of independent values of the probability which can be assigned. In the case of two constraints plus normalization (essentially three constraints) we could assign only three probabilities lying on the range 0 to 1 and then the other three would be uniquely determined.

Thus we see that for the white die there is still a statistically significant (at the 95% level) imperfection not explained by misplaced center of mass or oblateness. Jaynes 1979 says now that: "To assume a further very tiny imperfection [(the 2-3-6) corner chipped off] we could make even this discrepancy disappear; but in view of the (great) number of trials one will probably not consider the result as sufficiently strong evidence for this." The word "great" probably was intended to be "small".

Let us disagree midly with Jaynes at this point and actually look for this tiny third imperfection.

3. Three Constraints. Figure 3 gives a sketch of a die with the imperfection suggested by Jaynes.

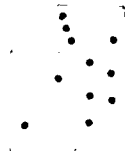

Figure 3. A die with a small chip broken off the 2, 3, 6 corner. Such an imperfection would tend to increase the probability of the die landing with the 2, 3, or 6 face showing "up".

By shifting its center of gravity, such a die would slightly favor the 2, 3 and 6 faces. Let us express this constraint as:

$$f_3(i) = -1, 1, 1, -1, -1, 1 \qquad (31)$$

Table V summarizes all three constraints we are now considering and attempts to simplify the algebra.

$$\text{Let} \quad w = \exp(-\lambda_3)$$

Table V Summary of the three constraints

i	$f_1(i)$	$f_2(i)$	$f_3(i)$	Contribution to Z	Factor out $x^{-2.5}y\,w^{-1}$
1	-2.5	1	-1	$x^{-2.5}y\,w^{-1}$	1
2	-1.5	1	1	$x^{-1.5}y\,w$	$x\,w^2$
3	-0.5	-2	1	$x^{-0.5}y^{-2}\,w$	$x^2\,y^{-3}\,w^2$
4	0.5	-2	-1	$x^{0.5}\,y^{-2}\,w^{-1}$	$x^3\,y^{-3}$
5	1.5	1	-1	$x^{1.5}\,y\,w^{-1}$	x^4
6	2.5	1	1	$x^{2.5}\,y\,w$	$x^5\,w^2$

Since the algebra gets a little tedious and mistakes are likely, the use of such a table is recommended in general. As a footnote, programs capable of simple algebra and differential calculus exist now. Use of such programs would be really beneficial. The three non-linear coupled equations for the constraints are now:

$$(2F_1+5) + (2F_1+3)\,x\,w^2 + (2F_1+1)\,x^2y^{-3}w^2 + (2F_1-1)x^3y^{-3} + (2F_1-3)x^4 + (2F_1-5)x^5\,w^2 = 0. \quad \textbf{(32)}$$

$$(F_2-1)\,(1+xw^2 + x^4 + x^5w^2) + (F_2+2)\,x^2\,y^{-3}\,(w^2+x) = 0.$$

$$(F_3+1)\,(1+x^3y^{-3}+x^4) + (F_3-1)\,w^2\,x\,(1+xy^{-3} + x^4) = 0.$$

With values: $F_1 = 0.0983$; $F_2 = 0.1393$; $F_3 = 0.0278$ the three coupled equations can be solved to give $x = 1.03072$; $y = 1.07425$; $w = 1.02159$ and $Z = 6.196106\,x^{-2.5}\,y\,w^{-1}$. Thus we get Table VI summarizing the resulting maximum entropy probabilities.

Table VI. Wolf's dice data with three constraints (white die)

i	g_i	p_i	$\Delta_i = g_i - p_i$	C_i
1	.16230	.16139	.00091	0.10
2	.17245	.17361	-.00116	0.16
3	.14485	.14434	.00051	0.04
4	.14205	.14256	-.00051	0.04
5	.18175	.18215	-.00040	0.02
6	.19660	.19594	.00066	0.04
				0.39

See footnote to Table II. Chi_c^2 on 2 degrees of freedom is 5.99.

The agreement between the observed frequencies g_i and the maximum entropy probabilities p_i is now essentially perfect. In fact it is too good! The agreement is much better than would be expected if Wolf's experiments had been repeated many times. The observed frequencies in many sets of experiments, each 20,000 tosses long, would differ from each other by much more than the $g_i - p_i$ from Table VI. Jaynes (1978) calculates that the fluctuations in the observed frequencies ought to be of order $(g_i/N)^{1/2}$. For $g_i = 1/6$, $\Delta g_i \sim 0.003$. All of the deviations $g_i - p_i$ in Table VI are smaller than this and all but $g_2 - p_2$ are about an order of magnitude smaller. Nevertheless, looking at Table IV again, with only two constraints, four of the deviations are larger than 0.003. In summary the observed frequencies for the white die can be completely explained by three physical constraints:

> The largest is No 2, the oblateness,
> The next largest is No 1, the center of gravity shift by spot removal and:
> The smallest is a tiny chip off the 2 - 3 - 6 corner.

The first two are required - the evidence for them is overwhelming. The evidence for the third is much weaker. From Table IV again, for two constraints, $Chi^2 = 9.37$ which is just significant at the 95% level but not significant at the 97.5% level.

Further thoughts on the white die. The computer program which solves the three constraint problem has been generalized (quite simply) to solve all of the imbedded problems:

> No constraints
> any one of the three acting by itself
> any two acting together
> all three.

The first case is trivial and reduces to $p_i = 1/6$. The last case has just been described. We summarize the results of all cases in Table VII.

Table VII. Chi squared for the white die. 1 = constraint on; 0 = off.

Constraints				
No. 1	No. 2	No. 3	df	Chi Square
1	1	1	2	0.39
1	1	0	3	9.37
0	1	1	3	56.28
0	1	0	4	72.01
1	0	1	3	189.77
1	0	0	4	199.42
0	0	1	4	253.85
0	0	0	5	270.96

In summary the most important single constraint is No. 2 (oblateness) the next important single is No. 1 center of gravity shift and the least important single is No. 3, corner chip. The best 2 constraints are 1 and 2 acting together followed by 2 and 3 and then 1 and 3. As a final footnote it is not sufficient to set one of the F's and its λ equal to zero and then solve the three equations. The equation for the inactive constraint must be dropped altogether and the corresponding λ set to zero. This has been done in the program.

b. The Red Die

To the best of my knowledge no one has ever attempted a complete analysis of the red die but with a simple program in place it becomes a trivial task to see if the same kind of thinking works just as well in this case. It had better! But we must be quite careful because although we expect similar kinds of asymmetries they need not be identical.

1. One Constraint. The first constraint as in the case of the white die, simply requires the average spot number. For the red die this value is: $\langle i \rangle = 3.49165$ which is less than 3.5. Even though this is less than 3.5 and not greater than 3.5 as expected we run the ME calculation with the one constraint:

$$\langle i-3.5 \rangle = -0.01835.$$

We get $x = 0.993728$ and $Z = 5.86966$. The ME probabilities are given in Table VIII.

Table VIII. Wolf's dice data with one constraint (red die)

i	g_i	p_i	$\Delta_i = g_i - p_i$	C_i
1	.17035	.16930	.00105	.13
2	.18155	.16824	.01331	21.07
3	.15880	.16718	-.00838	8.41
4	.14580	.16613	-.02033	49.77
5	.17240	.16509	.00731	6.47
6	.17110	.16406	.00704	6.05
				91.90

See footnotes to Table III.

Looking at $\Delta_i = g_i - p_i$ from Table VIII we see at once that Δ_3 and Δ_4 are negative while the others are all positive. This is precisely the same situation we found in Table III for the white die. The red die is also prolate in exactly the same way as the white die! This situation is not really as bizarre as might first be thought. Given that the die maker was prone to err on the prolate side, the only real coincidence is in the numbering of the faces. If he started his numbering (carving of spots) at the one spot he would be twice as likely to start with one of the four faces which are a short distance

apart as on either of the two "long" faces. Having done so, the two spots would be on short faces just as often as on a long face. Don't forget that once a one spot has been carved, the six must be on the opposite face. Thus the appearance of identical asymmetries on the two dice is not very surprising at all.

2. Two Constraints. We may now use the same program again to incorporate the first two constraints with values $F_1 = \langle f_1 \rangle = -0.01835$; $F_2 = \langle f_2 \rangle = 0.0862$. We get $x = 0.993965$; $y = 1.04508$; $Z = 5.66614^{2}x^{-2.52}y$ and Table IX gives the resulting probabilities.

Table IX Wolf's die data with two constraints (red die)

i	g_i	p_i	$\Delta_i = g_i - p_i$	C_i
1	.17035	.17649	-.00614	4.27
2	.18155	.17542	.00613	4.28
3	.15880	.15276	.00604	4.77
4	.14580	.15184	-.00604	4.80
5	.17240	.17227	.00013	0.00
6	.17110	.17123	-.00013	0.00
				18.13

See footnotes to Table III.

3. Three Constraints. We see here a tremendous improvement with an added bonus. Now that we have removed, by ME, the effects of the first two constraints, a third, smaller, but significant, constraint is now very obvious. Sides 5 and 6 have been fit very well indeed and the other four discrepancies are all of the same magnitude but with two plus signs and two minus signs. A possible physical explanation will be discussed later but the constraint to use now instead of the third constraint we used for the white die is:

$$f_3(i) = -1, 1, 1, -1, 0, 0 \qquad (33)$$

we now modify the master program slightly to accomodate this new constraint. Once again we can solve all of the imbedded problems. Table X shows the results.

Table X. Chi squared for the red die. 1 = constraint on: 0 = off.

Constraint 1	2	3	Degrees of freedom	Chi Square
1	1	1	2	0.08
0	1	1	3	2.40
1	1	0	3	18.13
0	1	0	4	20.44
1	0	1	3	74.86
0	0	1	4	77.16
1	0	0	4	91.90
0	0	0	5	94.19

Summarizing our results for the red die we have seen, that:

The red die was no more fair than the white die.

The excavation of spots and the subsequent shift of the center of gravity was not an important constraint for this die as it was for the white die. Other (unknown) compensatory constraints must have been at work.

The red die was oblate in essentially the same way that the white die was. For both dice this was the most important constraint.

There was no evidence of a corner chip here as there was for the white die but a constraint of the mathematical form -1, 1, 1, -1, 0, 0 was operating. No simple physical explanation seems in order but perhaps two simple constraints were acting in concert. A small wear spot on the 2 - 3 edge and a small excess of material on the 1 - 4 edge would make 2 and 3 more likely and 1 and 4 less likely.

After removing the most important constraint (oblateness) the misfit as expressed by $Chi^2 = 20.44$ is quite significant. Critical value Chi^2 on 4 df is 9.5 at 5% level.

When constraints number 2 and 3 are used together Chi^2 drops way down to 2.40 and the agreement between the observed frequencies g_i and the ME probabilities p_i is too good! Repetitions of the 20,000 toss experiment would very likely produce departures larger than the λ_i obtained from these two constraints.

The final conclusion from our exhaustive analysis of the two dice is that the maximum entropy principle allows us to discover physical imperfections in a pair of dice from data over 100 years old. At least as far as real dice are concerned, the principle of ME works and works brilliantly!

III. Published Criticisms

There have been many published papers which criticize the maximum entropy principle in general and Jayne's treatment of dice experiments in particular. Most of these attacks have been answered in the literature, some of them many times.

a. Older Criticism

For some of the earlier criticism see for example the paper by Rowlinson (1970) and Jaynes's (1978) answer. For a particularly virulent set of attacks see Friedman and Shimony (1971) and for defenses see Jaynes (1978) p 53, Tribus and Motroni (1972) Gage and Hestenes (1973) and Hobson (1972). See also Friedman (1973) and Shimony (1973) for their replies.

b. Frieden's Paper

The latest adventure in "anti-maximum-entropism" comes from B. Roy Frieden (1985) who professes to be "quite happy with (his) empirical results" using the maximum entropy formalism. The careful reader of Frieden's "Dice, Entropy and Likelihood" hereinafter referred to as DEL, might take pause at some of the statements to be quoted now.

Statement 1:

> "For example, this author originally believed ME to provide a maximum probable answer. However, at least for photon images, this is usually wrong. ..... Or, if it were required to estimate the most probable roll occurrences for an unknown die, the die would have to be known A priori to be fair, a rather restrictive assumption."

Wolf's dice were not fair. A priori, there is no requirement for fairness.

Statement 2:

> "Usually an engineer wants to know how probable his answer is, not how degenerate it is. The two concepts differ in general, and only coincide when every outcome has the same probability (i.e. when the die is fair."

The maximum-entropy die is fair only if there are no constraints acting besides normalization.

Statement 3:

> "The aim of this paper is to show that the die experiment just spoken of has solutions by classical, Bayesian estimation; that the probability of these solutions may be computed, as with any Bayesian problem; that therefore, there is no need to introduce a new

> concept such as maximum entropy in this most basic of problems; and that maximum entropy is not coincident with these solutions. In fact maximum entropy not only gives the wrong answer, it gives an answer that is very far from right."

Note the glee in the last sentence. Note also that the entire purpose of ME is to determine a prior probability assignment. This prior can then be used in any subsequent Bayesian analysis.

Statement 4:

> "we shall solve this problem in a purely classical way, without the need for recourse to any exotic estimator, such as ME."

Note the pejorative word "exotic".

Statement 5:

> "As we shall see, the most valid objection to the use of [Frieden's Eq.] (7) is that, although it describes 'maximum ignorance,' it does not describe the user's state for a die in particular. The wrong experiment is being performed to model maximum ignorance".

Frieden changes Jaynes' die problem brutally and then complains that his new problem is not the right problem.

Statement 6:

> "What this means is that we are not in a state of maximum ignorance when given an unknown die. We know what to expect a priori of its biases. For the particular case of a die, a real one, it would be wrong to assume maximum ignorance present. Hence, rolling a die is the wrong experiment to use when attempting to model 'maximum ignorance' situations. No wonder the result [Frieden's Eq.] (17) goes against intuition."

Once again, Frieden, having changed the problem, complains that this new problem is the wrong problem.

Statement 7:

> "We suggest that in the past readers have been seduced into a belief in ME principally because of this confusion between what constitutes maximum ignorance on one hand, and what constitutes the state of ignorance in a real die experiment on the other. If you want maximum ignorance do not consider a die experiment!"

Did you catch the truly pejorative word "seduced"?

Note in Statement 3, the use of the word "new" in connection with ME, and in Statement 4 the even more revealing word "exotic" which also appears again later. Note also the word "seduced" in Statement 7. A psychologist examining this paper might conclude that something other than pure scientific discourse is going on here. There is a pervasive feeling here that the author thinks he has found a fundamental flaw in the use of the ME principle and he is downright gleeful about it! Just reread Statement 3.

At this point we will examine the substance of the Frieden paper DEL.

Recall that in Jaynes' formulation of the problem, we are given:

> An enumeration of the possibilities,
>
> The average value of some linear constraint (e.g. the average spot values) measured in some previous experiment
>
> Normalization
>
> And nothing more.

In DEL, Frieden now changes the problem from that of a six sided real die to that of a three-sided imaginary die formed by combining rolls of one and six to yield one; two and five to yield two and three and four to yield three. He then calls the unknowns "biases" and labels them x_1, x_2, x_3. Then the real heart of the paper is introduced with Statement 8.

Statement 8:

> "By 'nothing' the user usually means that a priori every possible set of numbers x_1, x_2, x_3 (obeying normalization equation (1)) may be present with equal probability or frequency. Such a flat or uniform law is widely used in estimation problems. for example: when x_1, x_2, x_3 are the spatial coordinates of a material object whose location in a finite box is completely unknown a priori. Or, when a uniformly glowing planar image emits photons from unknown positions $(x,y) = x_1, x_2$. Or, when a distant aircraft of unknown coordinates (x,y) is being tracked; etc. This is also MacQueen and Marschak's (1975) definition of maximum ignorance, and we shall use it as well."

Here we go off the deep end! Frieden has changed an essentially discrete problem into an essentially continuous problem!

Recall the discussion in section Ia to the effect that Jaynes' die problem is isomorphic to any number of essentially discrete games, eg roulette, drawing a ball from a bag, drawing a card from a pack, etc. The essential features of these games are two in number: they are discrete and there is a symmetry principle operating. While small biases may be present in any of these games, large biases would be self defeating; they would be too easily detected. What the "user usually means" is, not only mathematically so vague as to be useless but also is completely irrelevent! Frieden can set up and attempt to solve any problem he choses. What he must not do is call his problem "Jaynes' problem"!

This Statement 8 changes Jaynes' problem by adding an enormous amount of information nowhere present in Jaynes' statement of the problem quoted above. Let us ask the question "how many bits would be required to encode the possible answers to Jaynes' problem"? Clearly for the three sided die, not even two bits would be necessary to encode the possible outcomes "1", "2" or "3". But if we are to take Statement 8 seriously we need another layer of information to discover which one of the infinite number of possible dice we are, in fact, shooting. Frieden, later in the paper, tries to simulate his continuous problem on a computer as follows:

Statement 9:

> "In other words, the prediction is that only roll outcomes 2 occurred! Actually this result can be explained in hindsight. Suppose we try to simulate the situation by repeatedly selecting sets of biases for a die, rolling the die, and only counting those biases which give rise to the required n. In this way, $p(x_1, x_2, x_3)$ is built up as a histogram, event by event. Let the biases be selected on a fine grid so that "every" triplet x_1, x_2, x_3 is sampled only once. This accomplishes the flat prior probability law [Frieden's Eq.] (7). Which such triplet will most often give rise to a value n = 2? It is obvious that the triplet (0,1,0) can only give rise to value n = 2."

Clearly B. Roy Frieden changed the problem - and drastically so. Frieden's problem now becomes: given an entire urn full of dice, all different, made very carefully by some imaginary machinist, so that each one will exhibit a different set of probabilities for the three faces. For a very crude set, with 11 possible probabilities for each face our patient die maker would manufacture 66 dice. Sixty-six is the number of normalizable triplets with a granularity of 0.1.

One real die for Jaynes, 66 imaginary dice for Frieden! And if Frieden wanted 101 possible probabilities for each face, our die maker would need to produce 5151 precisely carved dice! No wonder Frieden further changed the problem so that our old fashioned real six sided die lost half of its faces! Three - sided die indeed!

Now with our new three - sided die we are told that the average toss in a previous experiment was 2.0. Frieden now goes through some calculations to show that out of our urn containing a large number of dice, we have indeed selected the rare die with probabilities 0, 1, 0! Of course this screwball die would give an average toss of 2 - it had no choice. It had zero entropy - it always showed a 2 because it had to. Tossing this die yielded no new information, it couldn't. It was always pointless to toss it at all. What an enormous constraint to lower our entropy from a maximum to zero! Where in the original statement of the problem by Jaynes did it ever say that any face was impossible?

Frieden insists that his new problem represents a state of true ignorance and that the one single real Jaynes' die does not. We do not achieve a state of ignorance by making thousands of unnecessary assumptions! What we do is put in an enormous amount of prior information. Is it any wonder at all that Frieden's answer is wildly different from Jaynes?

Returning to the question asked about how many bits would be required for encoding the Frieden die, we see that we would first of all require $\log_2 (5151)$ or about 7 bits to encode the information "one die out of 5151 dice has been selected".

Let us examine Frieden's Monte Carlo calculation in a little more detail. If we use a granularity of 0.1 we will get 11 possible "biases" or probabilities for each face for a total of $(11)^3 = 1331$ dice. Of the 1331 dice only 66 can be normalized and of the 66 permissable dice only 6 will yield an expectation value of 2.0. These six have probabilities of (0,1,0), (.1,.8,.1) (.2,.6,.2), (.3,.4,.3), (.4,.2,.4), (.5,0,.5). The middle member of this set (.3,.4,.3) is the closest we can come to a "fair die" with probabilities (1/3,1/3,1/3).

For a granularity of 0.01, there will be 101 possible biases for each face (0., 0.01 ... 1.00). Thus there will be $(101)^3$ or 1,030,301 possible triplets, of which only 5151 can be normalized. From this set, any single choice will occur with probability 1/5151.

Of these 5151 dice only 51 would yield an expectation value of 2.0. These 51 would be (0.00,1.00,0.00), (0.01, 0.98, 0.01) ....(0.50, 0.00, 0.50). The closest to "fair" of any of these dice would be (.33, .34, .33).

Not only does Frieden change Jaynes' discrete problem into a continuous one to apply Bayes' Theorem, but he changes back to the discrete case when he "explains" Jaynes' ME approach. He says:

Statement 10:

> "Jaynes' ME approach [Frieden's refs] to the die problem is as follows. Assume that N is large enough [Frieden's Emphasis] that the law of large numbers [refs] holds, so that the die biases can be well approximated by values $g_i = n_i/N$.".

Did Frieden ever read Jaynes' paper? Where does Jaynes ever talk about N being large enough?

The only effect that N has is to determine the variance of the ME probabilities, not the probabilities themselves (p_i , i = 1,n). In fact in the same paper referenced by Frieden, Jaynes (1982) discusses an experiment with only N = 50 throws of a die in which we were given the average number of spots as 4.5 instead of 3.5 as expected from a fair die. Rowlinson (1970) advocated a binominal distribution instead of the ME distribution. We now quote Jaynes exactly: "Even if we come down to N = 50, we find the following. The sample numbers which agree most closely with (10, 16) while summing to $N_k = 50$ are $\{N_k\} = \{3,4,6,8,12,17\}$ and $\{N'_k\} = \{0,1,7,16,18,8\}$ respectively. With such small numbers, we no longer need asymptotic formulas. For every way in which Rowlinson's binominal distribution can be realized, there are exactly W/W' = (7!16!18!)/(3!4!6!12!17!) = 38,220 ways in which the maximum-entropy distribution $\{N_k\}$ can be realized". In the above statement, equations (10 and (16) are Jaynes' ME probabilities and Rowlinson's binominal probabilities respectively.

c. Musicus' Paper

The paper DEL by Frieden elicited a comment by Bruce Musicus (1986). Musicus accepted the Frieden transmogrification of Jaynes' discrete problem into the continuous problem we have already discussed. But Musicus made the excellent point that is nowhere mentioned in DEL that Frieden is discussing not probabilities but probability densities. Musicus proceeded to integrate Frieden's densities to generate marginal densities. With these marginal densities Musicus makes the point that no single point estimate would be at all useful or meaningful without a confidence region. Musicus then finds several "unreasonable" point estimates which he calls:

Statement 1:

$$\text{MAP - A: } x_1,x_2,x_3 = (0,1,0)$$

$$\text{MAP - B: } x_1,x_2,x_3 = \begin{cases} (0,0.5,0), & \text{for N even} \\ (0,0,0)\ \text{(sic)} & \text{for N odd} \end{cases}$$

We certainly agree with Musicus that these estimates are unreasonable.

Musicus adds:

Statement 2:

> "The fact that these point estimators all give radically different estimates is hardly surprising, given that the probability density in Frieden's problem is not unimodal, and is not strongly clustered around the center."

Musicus then proceeds to discuss Maximum Entropy as follows:

Statement 3:

> "Note that Maximum Entropy is thus justified for a problem involving known a priori biases x_1,x_2,x_3 and incomplete observation data (we only know the mean $\bar{n}$ of the throws of the dice, n_1,n_2,n_3) with asymptotically infinite numbers of throws N. Frieden's paper reverses the problem, asking for estimates of x_1,x_2,x_3 given the observation mean $\bar{n}$; it is not surprising that he gets a very different answer."

Fact: Using ME we are not given "a priori biases". It is the duty of the ME caluclation to convert information - the given mean $\bar{n}$ - into a probability distribution. No asymptotically infinite numbers of throws are necessary. Frieden's paper doesn't reverse the problem at all! Frieden changes an essentially discrete problem into an essentially continuous problem. We agree with Musicus' last statement "it is not surprising that Frieden gets a different answer".

d. Makhoul's Paper.

The Frieden paper we have been discussing was first pointed out to me at the Third ASSP Workshop on Spectrum Estimation and Modelling in a paper entitled "Maximum Confusion Spectral Analysis" by John Makhoul (1986). The content of this paper, which is available in the proceedings, was not quite as whimsical as its title suggested; at least two scientists in the audience seem to have been convinced by its attacks on the ME method, one of which was a simply a recounting of Frieden's paper. It was this presentation that stimulated me to study the subject of Jaynes' die in depth and ultimately to write this present paper. I am really indebted to John Makhoul for the stimulation. The Makhoul paper was limited in length to four pages of which only the first two are devoted to an "explanation" of ME and to the dice problem. The concentration of error per page in this paper is truly astounding!

Statement 1:

> "We assume that a random experiment has r possible events at each trial and that each event i, $1 \leq i \leq r$, has an a priori known probability x_i."

Fact: The prior probabilities are not known but unknown. The whole point of ME is to determine a set of prior probabilities consistent with all known information and maximally non-committal with respect to everything else!

Statement 2:

> "Perhaps the greatest contributing factor to the confusion surrounding ME is the claim or allusion by some that ME provides a posterior estimate of the a priori probabilities x."

Fact: ME is used to determine the prior probabilities. No competent ME practitioner, and certainly not Ed Jaynes, ever claims that ME produces posterior probabilities. As in the die experiment a sequence of ME calculations can produce sets of probabilities which agree better and better with observed frequencies, but each set of probabilities is essentially a prior probability assignment. If another experiment were then performed, Bayes equation would then use the ME probabilities and the experimental information to produce a set of posterior probabilities which might be better than the ME probabilities if the new information were neither redundant nor contradictory but cogent.

Statement 3:

> "Furthermore, it is claimed that this estimate is the most probable or most likely solution, ie, it is a maximum a posteriori (MAP) estimate. Also, it is claimed to be the solution that is 'maximally noncommittal' and makes the fewest assumptions in regard to the unknown data."

Fact: The first statement is untrue. The second is precisely correct, and the claim is also precisely correct.

Statement 4:

> "Far from being maximally noncommittal, the ME solution is based on a very specific and highly committal assumption of an equiprobable prior."

Fact: No equiprobable prior is ever claimed by competent ME practitioners. We have demonstrated in section Ib that under the assumption of discreteness (we have an enumeration of the

possibilities) and normalization and nothing more, equal probabilities for all possibilities is a consequence of ME, not an assumption. As soon as more information, perhaps in the form of expectation values, is provided, the ME probabilities become unequal in order to fit the observed constraints.

Statement 5:

> "The ME principle is then invoked to obtain the most likely vector of frequencies f that obey the constraint [Makhoul's Eq.] (10). Using our intepretation of the ME principle, we in effect assume that the die is a priori fair (unbiased) and then we compute the most likely frequencies for which (10) is true. If u = 4.5, which is very different from the expected value of 3.5 for a fair die, then the ME solution is given by [Makhoul's Eq.] (1)."

Fact: The primary goal of ME is to obtain a set of probabilities not frequencies. Ed Jaynes and other competent ME practitioners are always careful to distinguish between probabilities which can be assigned or calculated by ME or other valid procedures, and frequencies which can be measured in a laboratory. Under certain conditions which are elaborated in Jaynes (1968, 1978), there is a very strong correspondence between ME probabilities and measured frequencies but they are still quite distinct ideas conceptually. Once again the die is never assumed to be fair! Where does this gratuitous nonsense come from?

Statement 6:

> "While it is true that if N is large, having u = 4.5 is a good indicator that the die is most likely loaded because the probability of having u = 4.5 for a fair die is extremely small, the ME principle cannot be used productively to estimate the biases of the die. The ME die is simply not loaded. To name the problem the 'loaded die' problem has been a major source of confusion because it implies that the die is loaded and that the estimated frequencies are somehow related to the biases of the die. In ME, the die is known to be fair, but in an actual experiment the value of u comes out to be 4.5 for example instead of 3.5, which is a unlikely but possible event. We then use ME to compute the frequencies that most likely occurred from this most unlikely event."

Fact: N large (small, medium, known or unknown) is completely irrelevent for the solution of the ME problem! If N trials had been used to estimate frequencies then N would have a very large effect on the variance of the ME probabilities but none whatever on the probabilities themselves.

Fact: The straight jacket which says that ME die is not loaded is a complete fiction! It exists only in the mind of the author and has nothing to do with the theory and practice of ME methods. The reader is asked to refer again to the exhaustive analysis of the Wolf dice data. If this doesn't convince the reader that ME works beautifully to discover physical biases which were present in dice thrown repeatedly over 100 years ago, then nothing will.

The essential difficulty in Makhoul's paper in addition to his complete and total misunderstanding of ME, is his transformation, in agreement with Frieden and Musicus of our basically discrete dice problem into a strange unrecognizable continuous problem with objects which no one should ever call "dice".

ACKNOWLEDGEMENTS

The author is very pleased to acknowledge his gratitude to Elizabeth Galligan for her expert typing of so many drafts of this manuscript and to Theresa Walker of the AFGL Art Department for her very patient and expert setting of the many equations.

References

Cox. R., (1974). Probability, Frequency and Reasonable Expectation, Am. J. Physics, 17, 1.

Cox, R., (1961). The Algebra of Probable Inference, Johns Hopkins University Press, Baltimore, MD.

Czuber, E., (1908) Wahrscheinlichkeitsrechnung.

Hobson, A. (1972). The Interpretation of Inductive Probabilities, J. Stat. Phys 6, 189.

Frieden, B. Roy (1985). Dice Entropy and Likelihood, Proc. IEEE 73, 1764.

Friedman K., (1973), Replies to Tribus and Motroni and to Gage and Hestenes, J. Stat. Phys 9, 265.

Friedman K. and A. Shimony, (1971). Jaynes Maximum Entropy Prescription and Probability Theory, J. Stat. Phys. 3, 193.

Gage, D. W. and D. Hestenes, (1973). Comments on the paper "Jaynes Maximum Entropy Prescription and Probability Theory", J. Stat. Phys 7, 89.

Jaynes, E. T., (1957). Information Theory and Statistical Mechanics, Part I, Phys. Rev., 106, 620; Part II; ibid, 108,171.

Jaynes, E. T. (1963a), "Brandeis Lectures" in E. T. Jaynes Papers on Probability, Statistics and Statistical Physics, R. D. Rosenkrantz, Ed. D. Reidel Publishing Co., Boston, Mass.

Jaynes, E. T., (1968). "Prior Probabilities", IEEE Trans Syst. Sci Cybern., SSC4, 227.

Jaynes, E. T., (1978). Where do we stand on Maximum Entropy, in the Maximum Entropy Formalism, R. D. Levine and M. Tribus,Editors, MIT Press, Cambridge, Mass.

Jaynes, E. T., (1979). "Concentration of Distributions at Entropy Maxima" in E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics, R. D. Rosenkrntz, Ed., D. Reidel Publishing Co. Boston, Mass.

Jaynes E. T. (1982). "On The Rationale of Maximum - Entropy Methods", Proc. IEEE, 70 939.

Keynes, J.M., (1952). A treatise on Probability. MacMillam & Co, London.

deLaplace, Pierre Simon, (1951). A Philosphical Essay on Probabilities, Dover, New York.

Makhoul, J. (1986), "Maximum Confusion Spectral Analysis", Proc. Third ASSP Workshop on Spectrum Estimation and Modelling; Boston, Mass.

MacQueen J. and J. Marschak, (1975) "Partial Knowledge, Entropy and Estimation", Proc. Nat. Acad. Sci., Vol 72, pp. 3819-3824.

Rowlinson, J. S., (1970). Probability, Information and Entropy, Nature 225, 1196.

Shannon, C. E. and W. Weaver, (1949). The Mathematical Theory of Communication, The University of Illinois Press: Urbana.

Shimony, A., (1973). Comment on the interpretation of inductive probabilities, J. Stat. Phys 9, 187.

Teubners, B. G., Sammlung Von Lehr Buchern Auf Dem Gebiete Der Mathematischen Wissenschaften, Band IX p. 149, Berlin.

Tribus, Myron (1961), Thermostatics and Thermodynamics, D Van Nostrand Co., Princeton, N. J.

Tribus, Myron (1969), Rational Descriptions, Decisions and Designs. Pergamon Press, Oxford.

Tribus, Myron and H. Motroni, (1977) Comments on the Paper, Jaynes Maximum Entropy Prescription and Probability Theory", J. Stat. Phys 4, 227.